College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Structured documents and (X)HTML

Week 3

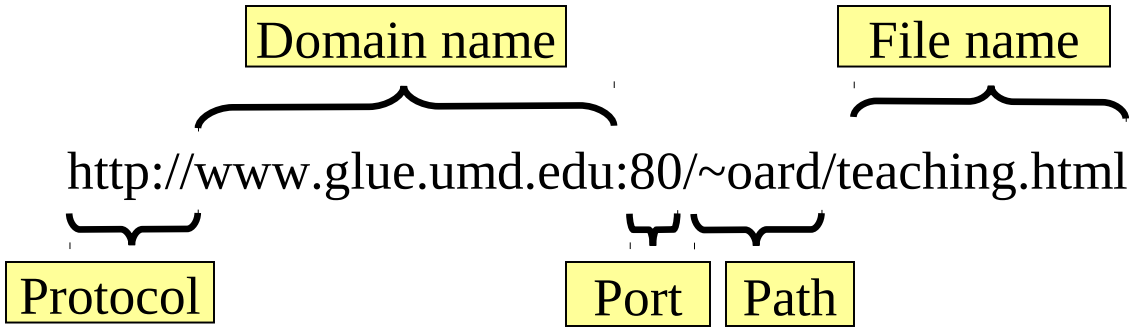
LBSC 690

Web Standards

- URL
  - Where to find the information
- HTTP
  - How to get the information
- HTML
  - How to write and interpret the information

Uniform Resource Locator (URL)

- Uniquely identify Web pages



- Request for directory retrieves “index.html”

HyperText Transfer Protocol (HTTP)

- Send request

  GET /path/file.html HTTP/1.0
  From: someuser@jmarshall.com
  User-Agent: HTTPTool/1.0

- Server response

  HTTP/1.0 200 OK
  Date: Fri, 31 Dec 2011 23:59:59 GMT
  Content-Type: text/html
  Content-Length: 1354
  <html><body> <h1>Happy New Year!</h1> … </body> </html>

HyperText Markup Language (HTML)

- Simple document structure language for Web
- Advantages
  - Adapts easily to different display capabilities
  - Widely available display software (browsers)
- Disadvantages
  - Does not directly control layout

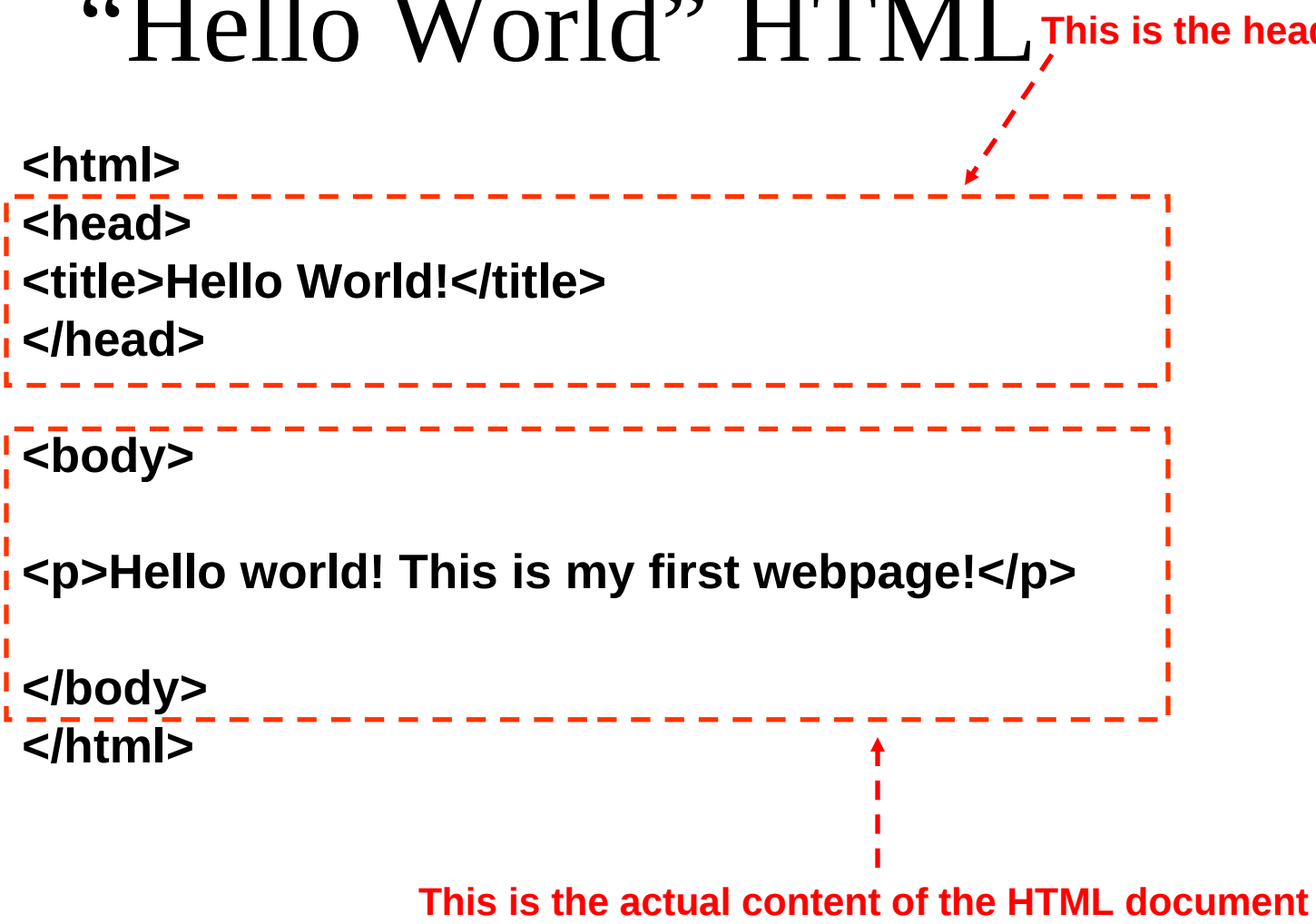
Rendering

- Different devices have different capabilities
  - Desktop
  - PDA
- Rendering maps logical tags to physical layout
  - Controls line wrap, size, font…
    - Place the title in the page border
    - Render <h1> as 24pt Times
    - Render as bold
- Somewhat browser-dependent
  - Slight differences in Internet Explorer, Firefox, Safari

HTML Document Structure

- “Tags” mark structure
  - <html>a document</html>
  - an ordered list
  - something in italics
- Tag name in angle brackets <>
  - Not case sensitive
- Open/Close pairs
  - Close tag is sometimes optional (if unambiguous)
- Element is tag and contents

“Hello World” HTML



Hands On:
Graphical Secure FTP

- Connect to “terpconnect.umd.edu”
  - If connecting for the first time, it asks about certificate, accept
- Change directory to “/pub/USERID”
- Upload or download files
- You can see these files at: http://terpconnect.umd.edu/~USERID/

HTML Page Editing Tips

- Edit files on your own machine
  - Upload when you’re happy
- Save early, save often, just save!
- Reload browser to see changes
- File naming
  - Don’t use spaces
  - Punctuation matters

Hands On:
Learning HTML From Examples

- Use browser to find a page you like
- On the “View” menu select “Source” (in IE9)
  - Opens a window with the source
- Compare HTML source with the Web page
  - Observe how each effect is achieved

Logical Structure Tags

- Head
  - Title
- Body
  - Headers: <h1> <h2> <h3> <h4> <h5>
  - Lists: <ol>, <ul> (can be nested)
  - Paragraphs:<p>
  - Tables: <table> <tr> <td> </td> </tr> </table>

Physical Structure Tags

- Appearance
  - Bold:
  - Italics:
- Font tag
  - Typeface:
  - Size:
  - Color:

(Hyper)Links

Source Document: index.html

```
<html>
<head>
<title>Hello World!</title>
</head>
<body>
<p>Hello world! This is my first webpage!</p>
<p>Click <a href="test.html">here</a> for another page.</p>
</body>
</html>
```

Target Document: test.html

```
<html>
<head>
<title>Another page</title>
</head>
<body>
<p>This is another page.</p>
</body>
</html>
```

Hypertext “Anchors”

- **Internal anchors**: somewhere on the same page
  - Students
    - Links to: Student Information

- **External anchors**: to another page
  - iSchool
  -
    Douglas W. Oard

Paths in Anchors

- URL may be complete or relative to current page
- In: ischool.umd.edu/content/prospective-students
  - url: href="http://ischool.umd.edu/content/research"
  - equivalent to relative path: href="/content/research"
  - and even this: href="research"

- File name part of URL is case sensitive (on Unix servers)
  - Protocol and domain name are not case sensitive

Images

- <img src="*URL*"> *or* <img src="*path/file*">
  - <img src="http://www.ischool.umd.edu/sites/all/themes/ischool/images/ischool_logo_home.gif">
  - At http://www.ischool.umd.edu, equivalent to:
  - <img src="/sites/all/themes/ischool/images/ischool_logo_home.gif">
  - SRC: can be url or path/file
  - ALT: a text string if not showing image

Image Hyperlinks

- Can use image within anchor to link:

  <img src=*IMAGEURL*>

  Example:

  <img
  src="/sites/all/themes/ischool/images/um_logo.jpg"
  alt="University of Maryland" />

Tables

<table>

<tr><td> eenie </td><td>meenie</td><td> miney</td> </tr>

<tr><td> mo </td><td> catch </td><td> a tiger</td> </tr>

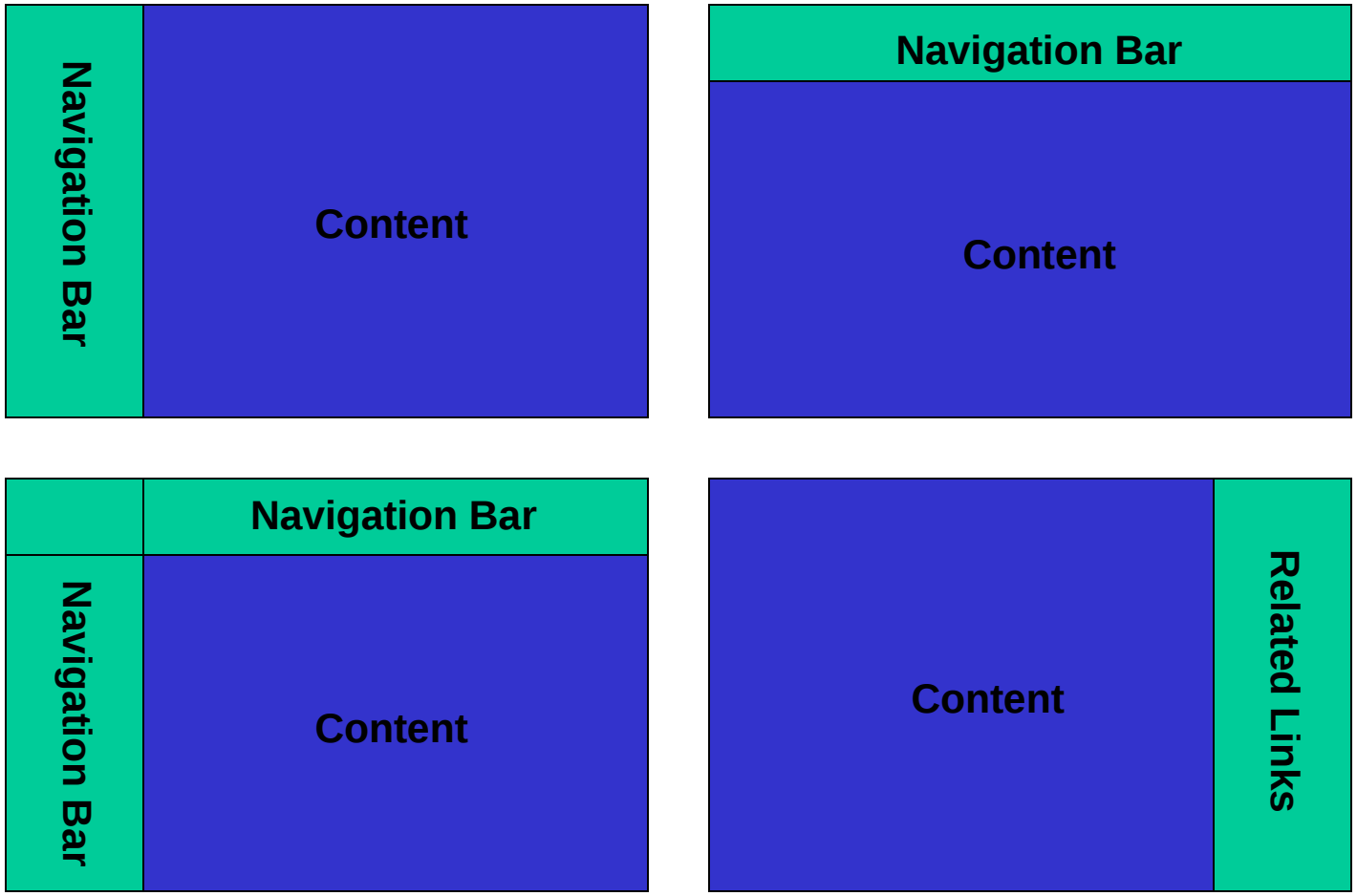
<tr><td> by </td><td> the </td><td> toe </td> </tr>

</table>

Table Example

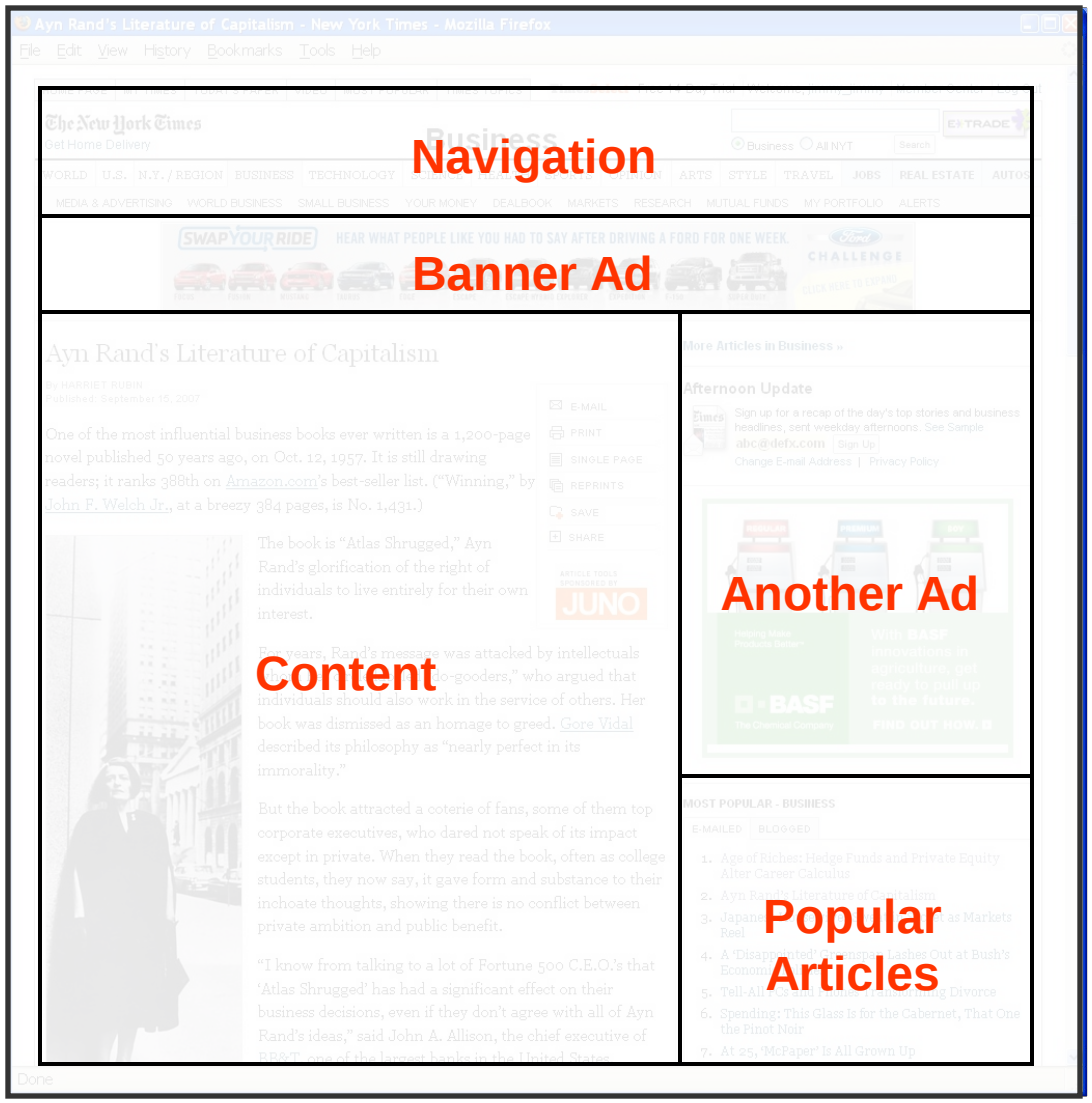
```
<table align=“center”>
<caption align=“right”>The caption</caption>
   < tr align=“LEFT”>
        <th> Header1 </th>
        <th> Header2</th>
   </tr>
   <tr><td>first row, first item </td>
        <td>first row, second item</td></tr>
   < tr><td>second row, first item</td>
        <td>second row, second item</td></tr>
</table>
```

Common HTML Pattern: Grid Layouts

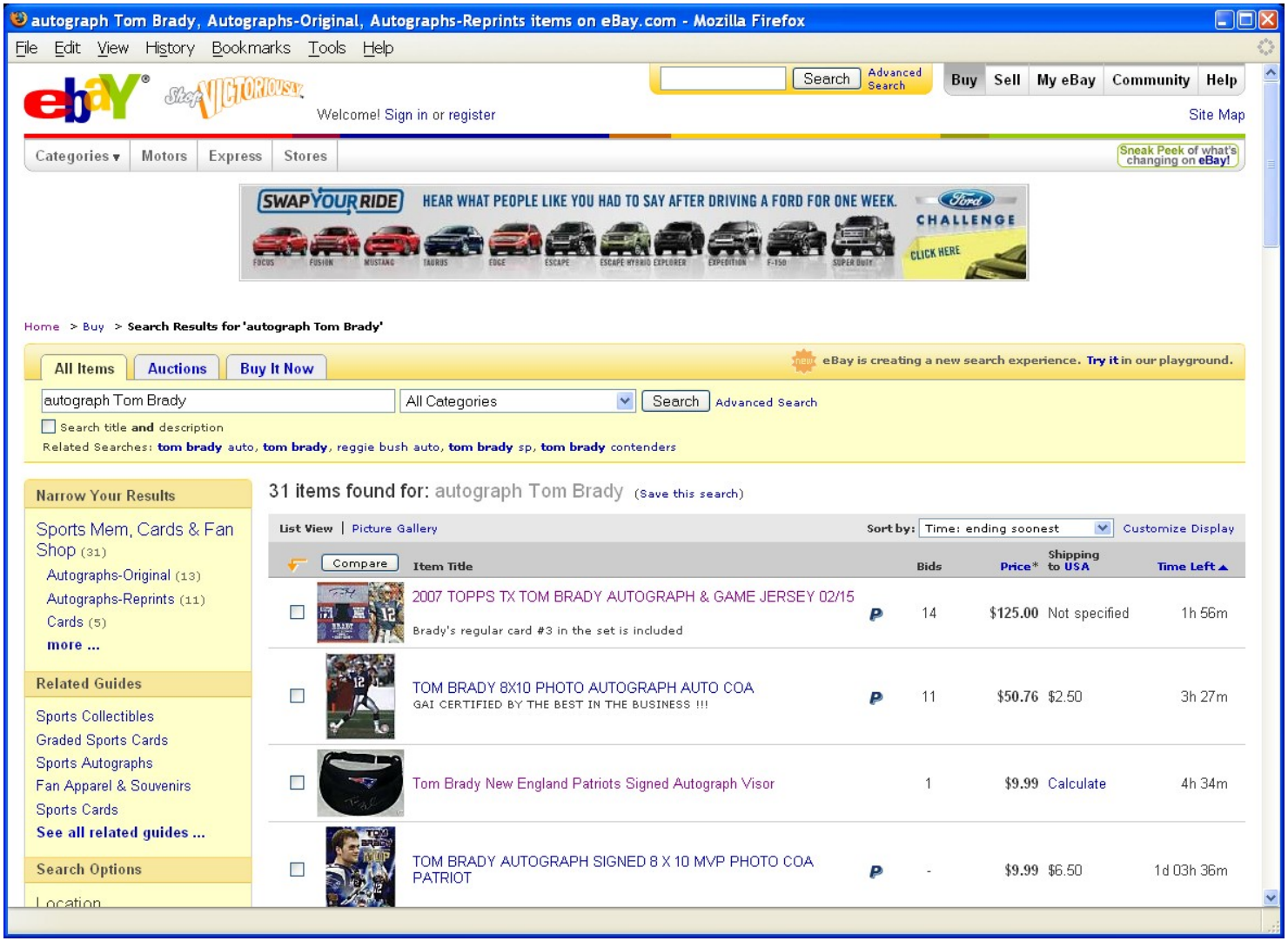


Grid Layout: NY Times

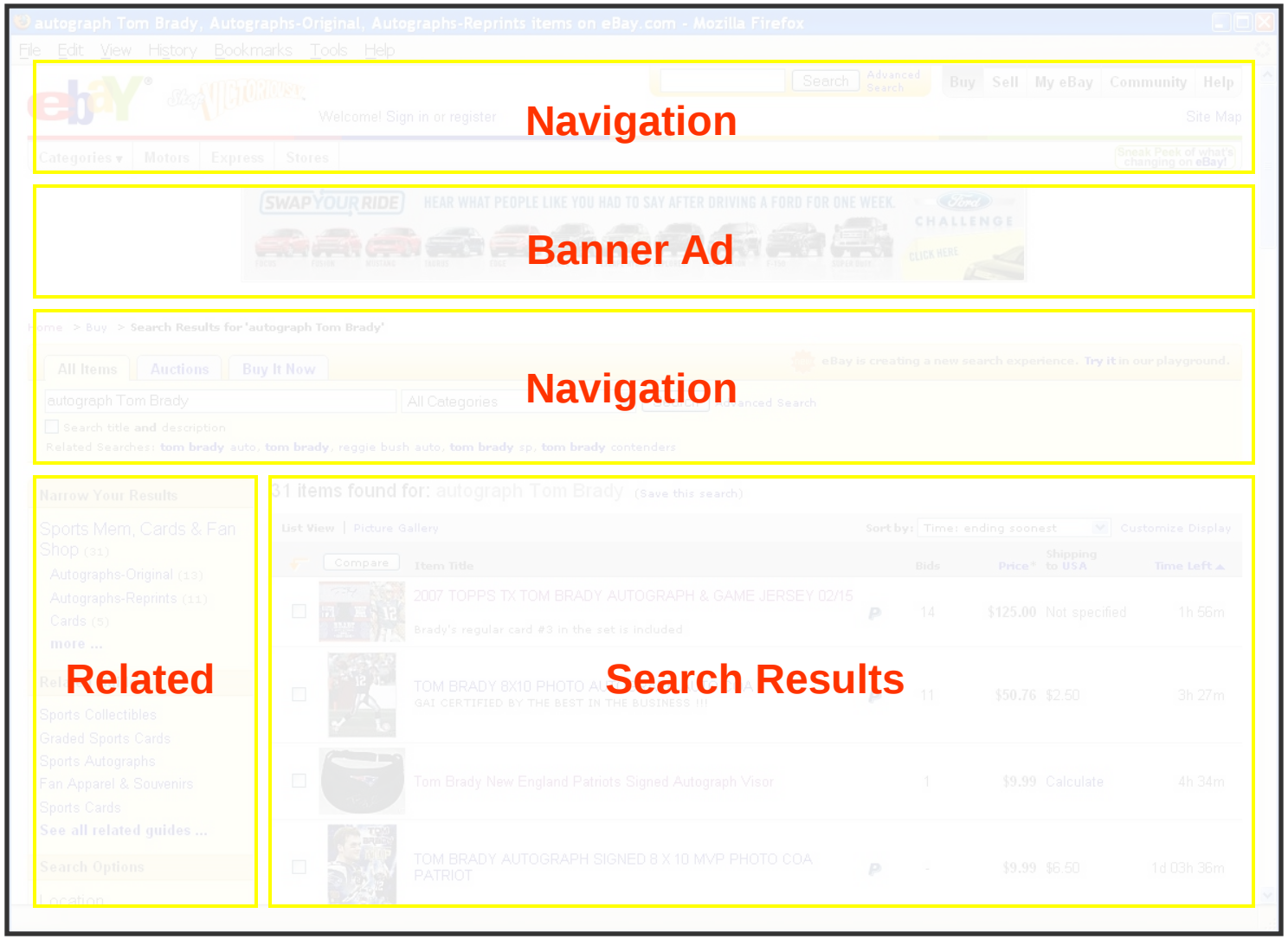
Grid Layout: NY Times



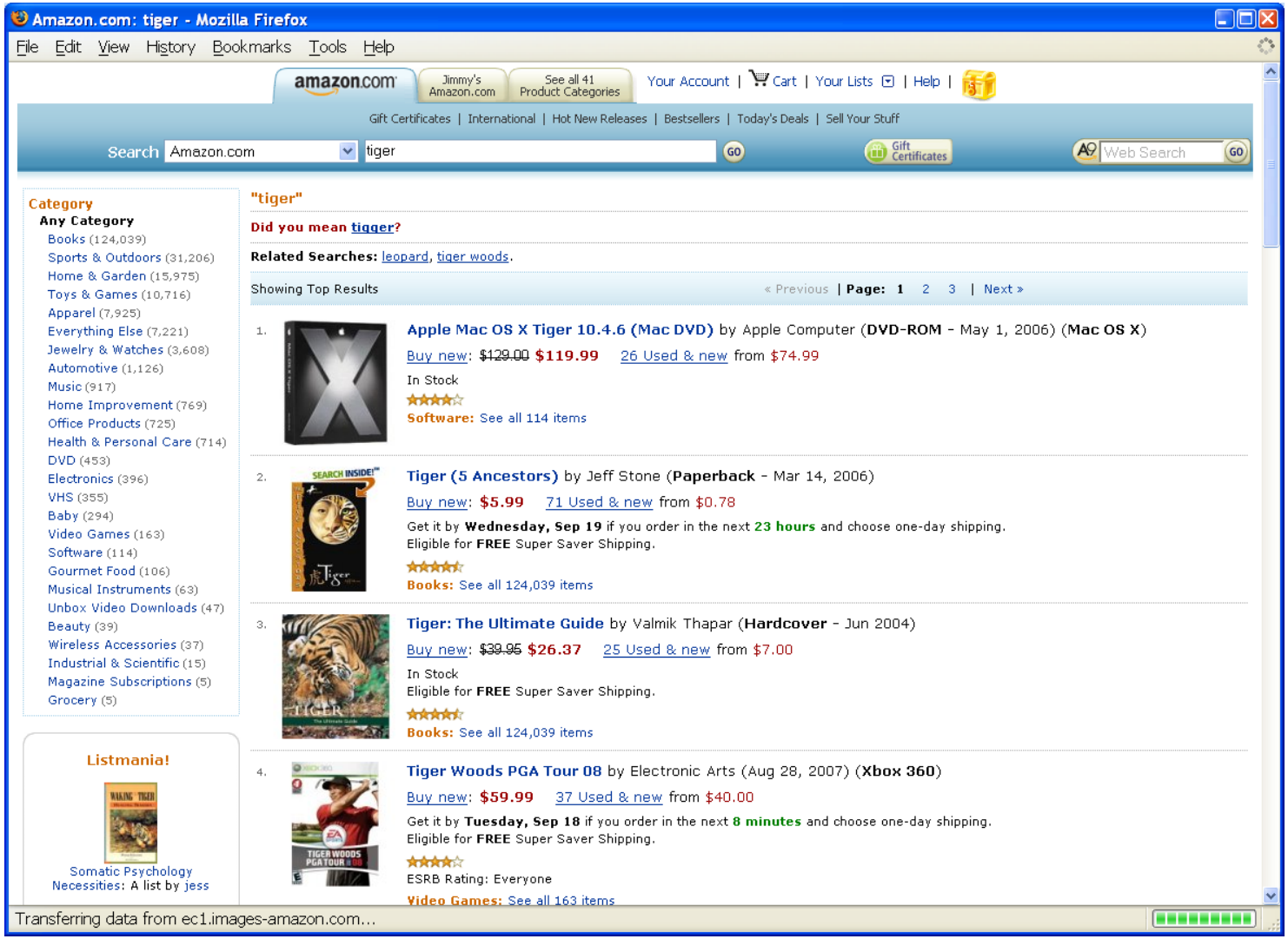
Grid Layout: ebay



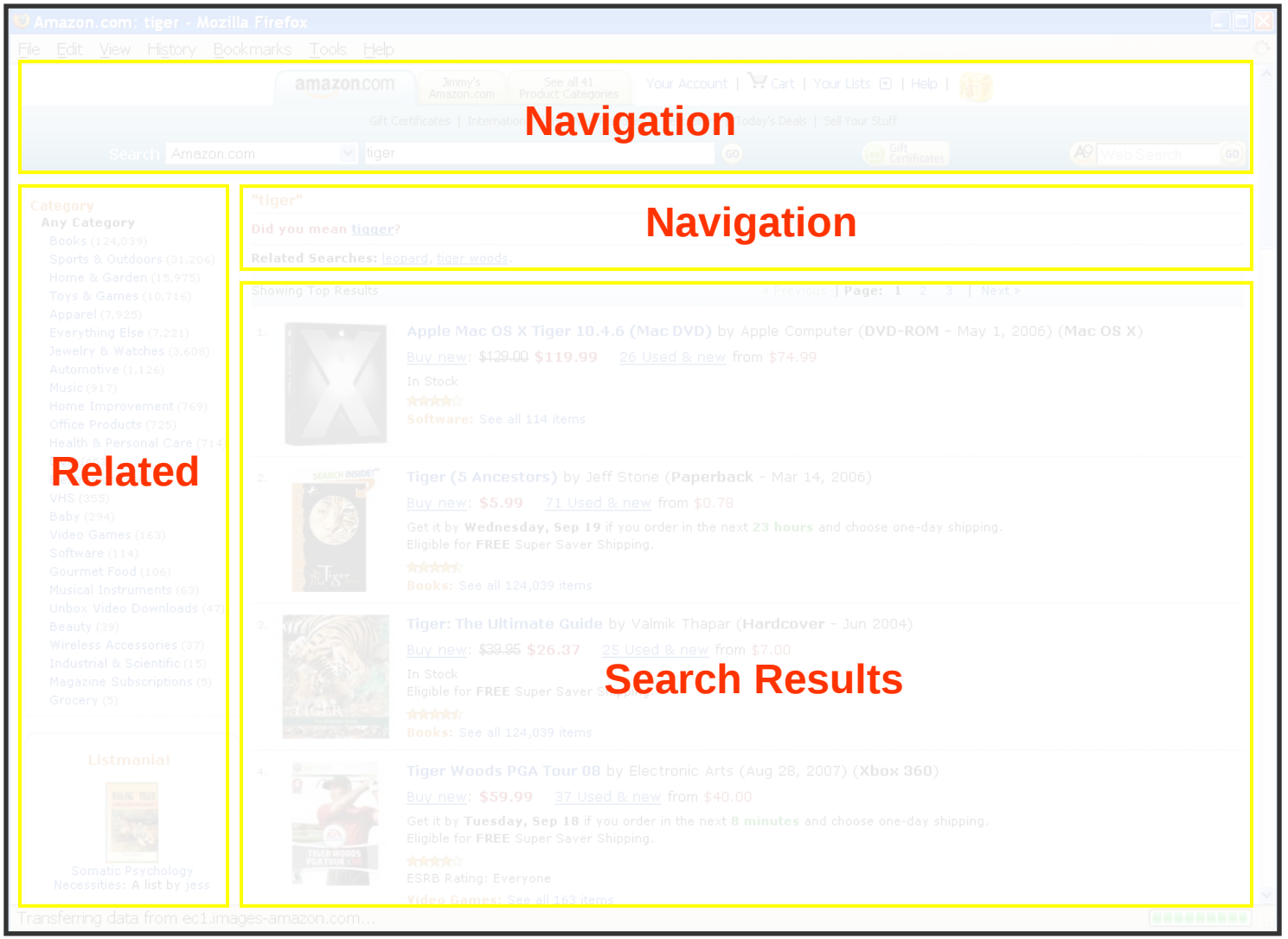
Grid Layout: ebay



Grid Layout: Amazon



Grid Layout: Amazon



Markup Languages: Writing Wikipedia Pages

- Wikipedia uses MediaWiki software
- MediaWiki has its own markup language
  - http://en.wikipedia.org/wiki/MediaWiki
  - MediaWiki syntax vs. Equivalent HTML and Rendered output

Wikipedia Infobox

- Infoboxes summarize important points
  - Types for business, biological taxonomy, video game characters, etc.
- Simple markup language
- Expanded to generate table for infobox

Structured Documents: Wikipedia Page Organization

- Table of Contents
  - We just enter the section titles
  - MediaWiki creates TOC box with internal links
- References
  - We just put citations within text
  - MediaWiki numbers them and creates a section

WYSIWYG

- Anything you can do you can see
  - Everything you can do you can see
- MediaWiki markup helpful for pages
- Need HTML for e.g. representing forms

HTML Editors

- Several are available
  - Adobe Dreamweaver available commercially
  - Microsoft Word
- Tend to use physical layout tags extensively
  - Detailed control can make hand-editing difficult
- You may still need to edit the HTML file
  - Some editors use browser-specific features
  - Some HTML features may be unavailable
  - File names may be butchered when you upload

HTML Validators

- Syntax checking: cross-browser compatibility
  - http://validator.w3.org
  - Try it on http://www.umd.edu ☹
- Style checking: Section 508 compliance
  - http://www.cynthiasays.com/
  - Try it on http://www.umd.edu ☹

What's Wrong with the Web?

- HTML
  - Confounds structure and appearance (XML)
- HTTP
  - Can't recognize related transactions (Cookies)
- URL
  - Links breaks when you move a file (PURL)

What’s a Document?

- Content
- Structure
- Appearance
- Behavior

History of Structured Documents

- Early standards were “typesetting languages”
  - NROFF, TeX, LaTeX, SGML
- HTML was developed for the Web
  - Too specialized for other uses
- Specialized standards met other needs
  - Change tracking in Word, annotating manuscripts, …
- XML seeks to unify these threads
  - One standard format for printing, viewing, processing

eXtensible Markup Language (XML)

- SGML was too complex
- HTML was too simple
- Goals for XML
  - Easily adapted to specific tasks
    - Rendering Web pages
    - Encoding metadata
    - “Semantic Web”
  - Easily created
  - Easily processed
  - Easily read
  - Concise

Some Basic Rules for XML

- XML is case sensitive
- XML declaration is the first statement
  - <?xml version="1.0"?>
- An XML document is a “tree”
  - Must contain one root element
  - Other elements must be properly nested
- **All** start tags must have end tags – ok if same
- Attribute values must have quotation marks
  - <item id=“33905”>
- Certain characters are “reserved”
  - For example: < is used to represent <

Really Simple Syndication (RSS)

```
<?xml version="1.0"?>
<rss version="2.0">
<channel>
   <title>Lift Off News</title>
   <link>http://liftoff.msfc.nasa.gov/</link>
   <description>Liftoff to Space Exploration.</description>
   <language>en-us</language>
   <pubDate>Tue, 10 Jun 2003 04:00:00 GMT</pubDate>
   <lastBuildDate>Tue, 10 Jun 2003 09:41:01 GMT</lastBuildDate>
   <docs>http://blogs.law.harvard.edu/tech/rss</docs>
   <generator>Weblog Editor 2.0</generator>
   <managingEditor>editor@example.com</managingEditor>
   <webMaster>webmaster@example.com</webMaster>
   <ttl>5</ttl>
   <item>
      <title>Star City</title>
      <link>http://liftoff.msfc.nasa.gov/news/2003/news-starcity.asp</link>
      <description>How do Americans get ready to work with Russians aboard the International Space Station? They take
                  a crash course in culture, language and protocol at Russia's Star City.</description>
      <pubDate>Tue, 03 Jun 2003 09:39:21 GMT</pubDate>
      <guid>http://liftoff.msfc.nasa.gov/2003/06/03.html#item573</guid>
   </item>
</channel>
</rss>
```

See example at http://www.nytimes.com/services/xml/rss/

XML: A Family of Standards

- Definition: DTD
  - Known types of entities with “labels”
  - Defines part-whole and is-a relationships
- Markup: XML
  - “Tags” regions of text with labels

Document Type Definition (DTD)

```
<!ELEMENT poem ( (title, author, stanza)* )
>
 <!ELEMENT title (#PCDATA) >
<!ELEMENT author (firstname, lastname) >
<!ELEMENT firstname (#PCDATA) >
 <!ELEMENT lastname (#PCDATA) >
 <!ELEMENT stanza (line+ | linein+) >
 <!ELEMENT line (#PCDATA) >
<!ELEMENT linein (#PCDATA) >
```

#PCDATA	span of text
a,*b*	*a* followed by *b*
a|*b*	either *a* or *b*
*a**	0 or more *a*'s
a+	1 or more *a*'s

XHTML Example

- View “The Song of the Wandering Aengus”
  - http://www.umiacs.umd.edu/~oard/teaching/690/spring08/notes/3/xml.htm
- Built from three files
  - yeats01.xml
  - poem01.dtd
  - poem01.xsl

XML Example

```
<?xml version="1.0"?>
<!DOCTYPE POEM SYSTEM "poem01.dtd">
<?xml-stylesheet type="text/xsl" href="poem01.xsl"?>
<POEM>
        <TITLE>The Song of Wandering Aengus</TITLE>
        <AUTHOR> <FIRSTNAME>W.B.</FIRSTNAME>
                        <LASTNAME>Yeats</LASTNAME>
        </AUTHOR>
<STANZA>
        <LINE>I went on to the hazel wood,</LINE>
        <LINEIN>Because a fire was in my head,</LINEIN>
         <LINE>And cut and peeled a hazel wand,</LINE>
</STANZA>
</POEM>
```

XHTML: Writing HTML as XML

```
<?xml version="1.0" encoding="iso-8859-1"?>
<html xmlns="http://www.w3.org/TR/xhtml1" >
<head>
   <title> Title of text XHTML Document </title>
</head>
<body>
<div class="myDiv">
    <h1> Heading of Page </h1>
     <p> here is a paragraph of text. I will include  inside this paragraph
           a bunch of wonky text so that it looks fancy. </p>
     <p>Here is another paragraph with  <em>inline emphasized</em>
           text, and <b> absolutely no</b> sense of humor. </p>
     <p>And another paragraph, this one  with an <img src="image.gif"
          alt="waste of time" />   image, and a <br /> line break. </p>
</div>
</body></html>
```

Even More Uses of XML …

- **CML –** Chemical Markup Language
- **CellML –** biological models
- **BSML –** bioinformatic sequences
- **MAGE-ML –** MicroArray Gene Expression
- **XSTAR –** for archaeological research
- **MARCXML –** MARC in XML
- **AML –** astronomy markup language
- **SportsML –** for sharing sports data

Some Layout Guidelines

- **C**ontrast: make different things different
  - to bring out dominant elements
  - to create dynamism
- **R**epetition: reuse design throughout the interface
  - to achieve consistency
- **A**lignment: visually connect elements
  - to create flow
- **P**roximity: make effective use of spacing
  - to group related and separate unrelated elements

Summary

- Learning to build simple Web pages is easy
  - Which is good news for the homework!
- Grid Layout facilitated by tables and CSS
- All documents are structured documents
  - But some expose the structure better than others
- XML is a flexible markup language
  - Complete separation of structure and appearance

Before You Go

On a sheet of paper, answer the following (ungraded) question (no names, please):

What was the muddiest point in today's class?